Evolutionary Explanations of Distributive Justice

J. McKenzie Alexander

Philosophy of Science, Vol. 67, No. 3 (Sep., 2000), 490-516.

Stable URL:
http://links.jstor.org/sici?sici=0031-8248%28200009%2967%3A3%3C490%3AEEODJ%3E2.0.CO%3B2-%23

Evolutionary Explanations of Distributive Justice*

J. McKenzie Alexander†‡
Logic & Philosophy of Science, University of California, Irvine

Evolutionary game theoretic accounts of justice attempt to explain our willingness to follow certain principles of justice by appealing to robustness properties possessed by those principles. Skyrms (1996) offers one sketch of how such an account might go for divide-the-dollar, the simplest version of the Nash bargaining game, using the replicator dynamics of Taylor and Jonker (1978). In a recent article, D'Arms et al. (1998) criticize his account and describe a model which, they allege, undermines his theory. I sketch a theory of evolutionary explanations of justice which avoids their methodological criticisms, and develop a spatial model of divide-the-dollar with more robust convergence properties than the models of Skyrms (1996) and D'Arms et al. (1998).

1. Introduction. In a recent article, D'Arms, Batterman, and Górny (1998) examine the evolutionary game theoretic account of justice suggested by Skyrms (1996). In their discussion, they contrast Skyrms's explanatory strategy with that favored by contemporary evolutionary psychologists (these are the methods of evolutionary generalism and evolutionary particularism, respectively). They offer three criteria for evaluating evolutionary accounts of moral norms (representativeness, robustness, and flexibility), and argue that Skyrms's account fares less well than one might hope for with respect to representativeness and robustness.

Although I agree with much of what D'Arms et al. have to say, this paper challenges their conception of the structure of evolutionary explanation and describes an evolutionary model with very robust convergence

*Received September 20, 1999. Revised April 19, 2000.

†Send requests for reprints to the author, Logic & Philosophy of Science, School of Social Sciences, 3151 Social Sciences Plaza, University of California, Irvine, CA 92697-5100; email: jalex@uci.edu.

‡I would like to thank Brian Skyrms, Penelope Maddy, and two anonymous referees for their helpful comments and suggestions regarding an earlier draft of this paper.

Philosophy of Science, 67 (September 2000) pp. 490–516. 0031-8248/2000/6703-0008$2.00

properties. In Section 3, I offer an alternative view of the structure of evolutionary explanation, one in which evolutionary generalism and evolutionary particularism are not seen as competing strategies but as complementary components of a single schema. I argue that this conception not only avoids some of the criticisms leveled against Skyrms, but has, as an additional advantage, the possibility of explaining some of the normative aspects of justice. In Section 4, I develop further my spatial evolutionary model discussed in Alexander and Skyrms (1999), and present results supporting the evolutionary game theoretic explanation of justice.

2. Evolutionary Models of Distributive Justice. As detailed discussions of Skyrms's model can be found elsewhere, I shall be brief in my reconstruction, primarily emphasizing the points at which Skyrms (1996) and D'Arms et al. (1998) differ. Both use a simplified version of the bargaining game discussed in Nash (1950); in this game, two players must divide a good (say a cake) sliced into N pieces. Each player decides how much of the cake she wants, in terms of the number of slices, without communicating her choice to the other. If the individual demands do not sum to more than the total number of slices, each player gets what he or she desired. If the sum of individual demands exceeds the total number of slices, each player receives nothing.

Skyrms uses the discrete replicator dynamics of Taylor and Jonker (1978) to model a population of agents. According to this model, the state of the population at a particular time t is represented by a vector $\langle p_0,\ldots,p_N\rangle$, where $p_i \in [0,1]$ represents the proportion of the population desiring i slices of the cake. In the next generation, the population proportion changes according to the formula

$$p_i' = p_i + (e_i + b)/(P + b) \tag{1}$$

where e_i denotes the expected fitness of strategy i in the population, P the average fitness of the population, and b the background fitness of the population.

Representing the population this way assumes an infinite population of agents. D'Arms et al. (1998) relax this assumption and assume only a finite population of agents who randomly interact with each other. "Random interaction," in this context, means pairwise sampling without replacement until each agent in the population has interacted with someone. At the beginning of the next generation, the population is renormalized to keep the total number of agents constant, allocating strategies as determined by the payoffs in the previous round.

Given the different assumptions underlying each model, one might expect them to produce divergent results; surprisingly, this does not happen. Plots of population trajectories over the simplex space, for the special case

where the set of possible strategies is restricted to demand 1/3, demand 1/2, and demand 2/3, look virtually identical. Both contain a significant region that converges to a state of fair division with another, smaller, region converging to a polymorphism between demand 1/3 and demand 2/3.[1]

The existence of regions converging to unfair polymorphisms like the 1/3–2/3 split poses a problem for evolutionary game theoretic accounts of justice. When the basin of attraction for that region is of a significant size, the strength of the evolutionary explanation is weakened because of the dependence on the initial conditions. Saying that we opt for fair division in completely symmetric circumstances because the initial state S of our society happened to fall within the basin of attraction of fair division, at best, offers only a weak inductive-statistical explanation.

Skyrms addresses this by showing how introducing small amounts of correlation between strategies changes the size of the basins of attraction. When self-correlation between strategies exceeds .2, the basins of attraction for the unfair polymorphisms virtually disappear. This would, it seems, complete the evolutionary explanation: we have shown how from every (or almost every) initial state of the population the evolutionary dynamics carry the population to a final state where fair division dominates.

Unfortunately, there remains the question of what allows us to include a small amount of self-correlation in the model. Some plausible explanation must be given to prevent the modification from appearing ad hoc. According to D'Arms et al., Skyrms's commitment to the explanatory schema of evolutionary generalism creates a problem for him on precisely this point. Evolutionary generalism, which requires one remain silent on the question of what proximate mechanisms might account for the behavior of fair division, severely limits the possible explanations one can give to justify introducing self-correlation into the model. They note, "for Skyrms to suggest

1. Strictly speaking, the nature of D'Arms et al.'s model prevents us from speaking of regions of the simplex space converging to fair division in the same sense as with Skyrms's model. The fact that D'Arms et al. use a finite population with random pairing introduces a stochastic element into their model which prevents us from being able to say with absolute certainty that the population will always follow a certain trajectory when started at a particular point in the simplex space. To see this, notice that certain odd trajectories may occur in their model which cannot occur under the replicator dynamics. For example, in a population containing only the strategies demand 2/3 and demand 1/3, random pairing could, conceivably, pair all agents who demand 2/3 with agents who demand 2/3. If this continued, the demand 2/3 strategy could eventually go extinct. This cannot happen in the replicator dynamics. Thus, with respect to D'Arms et al.'s model, one can only say that there exist certain regions of the simplex space such that any population started in those regions will, with high probability, converge to a polymorphism. Similar qualifications need be made for regions which "converge" to fair division.

that we do [have genetic proclivities for strategies in the Nash demand game] would involve an uncomfortable amalgam of generalist and particularist explanatory schemas: insisting on an innate biological disposition toward a strategy without offering any concrete account of or evidence for the psychological mechanisms that subserve it." (D'Arms et al. 1998, 92) I shall return to the question of whether this composition of explanatory schemas is, in fact, an "uncomfortable amalgam."

Setting this objection aside, for the moment, let us assume that one can justify including a certain amount of correlation into the model. D'Arms et al. observe that if we are going to introduce correlation into the model, it cuts both ways: strategies can be both positively and negatively correlated. Although it pays for the demand ½ strategy to self-correlate, since both agents benefit from the interaction, it would be foolish for the demand ⅔ strategy to self-correlate as both agents receive nothing. (Note that introducing correlation, whether positive or negative, has absolutely no effect on the demand ⅓ strategy since this strategy always receives ⅓ no matter with whom it pairs.) In the finite population model of D'Arms et al, when positive self-correlation applies only to demand ½, the basin of attraction for the unfair polymorphism fails to disappear; introducing anticorrelation between demand ⅔ strategies causes the basins of attraction for the unfair polymorphism to grow. D'Arms et al. interpret this as undermining Skyrms's account of the evolution of justice. This conclusion seems to me too strong, as I shall explain.

3. The Structure Of Evolutionary Explanations. To begin, a few clarificatory remarks, concerning what the explanatory schemas of evolutionary particularism and evolutionary generalism are, seem in order. Unfortunately, D'Arms et al. do not explicitly characterize the explanatory schemas of evolutionary generalism and evolutionary particularism, describing instances of these explanatory schemas instead. D'Arms et al. begin their discussion of evolutionary particularism by focusing on particular approaches to the evolutionary explanation of human behavior, later narrowing it to the case where moral capacities are the explanandum (presumably in anticipation of the later discussion regarding fair division in the Nash bargaining game). This is suggested, somewhat elliptically, at the beginning of their discussion of evolutionary particularism, for they write: "According to the particularist hypothesis, the human mind comprises an array of discrete adaptive mechanisms . . . " (D'Arms et al. 1998, 84) If this is to be a description of the explanatory schema of evolutionary particularism, then evolutionary particularism *only* serves to explain features of the human mind. However, it seems to me that the explanatory schemas of evolutionary particularism and evolutionary generalism have

a wider range of potential applicability than this would allow. Thus, to arrive at a characterization of the explanatory schema we must work backward from the particular instances provided.

According to D'Arms et al., the schema of evolutionary particularism, when applied in the domain of evolutionary psychology, generates explanations of the following sort:

> the human mind comprises an array of discrete adaptive mechanisms, generated through a process of natural selection in which distinctive sorts of adaptive problems forged functionally distinct adaptive solutions . . . These mechanisms are functionally specialized to process information concerning specific adaptive problems and produce behavior that solves those problems. (D'Arms et al. 1998, 84)

I interpret the statement that "distinctive sorts of adaptive problems forged functionally distinct adaptive solutions" as saying that, for each adaptive problem p, natural selection generates a mechanism M_p such that M_p extracts information particular to the problem p from the current situation S (D'Arms et al. call this the "environment of evolutionary adaptation"), producing a final behavior b that solves p.

When the explanandum is a moral capacity, D'Arms et al. characterize evolutionary particularism as follows:

> Thus, for instance, the particular hypothesis with respect to our moral capacities holds that selective pressures deriving from the fitness consequences of various social relations . . . have forged similarly specific adaptive psychological mechanisms which mediate cognition and motivation in these domains. (D'Arms et al. 1998, 82)

I interpret this as saying that a particularist explanation of a (particular) moral capacity c consists of specifying an adaptive mechanism M_c which serves to extract information from the current situation S, thus "mediat[ing] cognition," and motivating the appropriate behavior b. D'Arms et al. presumably slip from speaking of the mechanism *producing* the behavior to merely *motivating* the behavior since, in the case of the moral phenomena, the "morally correct" behavior is not always produced.

In light of this, I take the explanatory schema of evolutionary particularism to have the following form. The explanandum consists of a behavior b in a situation S in response to problem p. The explanans consists of a mechanism M_p which, given p in S, produces b. Since the mechanism M_p is tailor-made to generate the behavior b in response to the adaptive problem p, the name "evolutionary particularism" seems apt. Thus, an evolutionary particularist explanation of, say, some individual's behavior in the Nash bargaining game, would consist of a specification of some

mechanism *M* that serves to produce the observed behavior in the appropriate circumstances.

Now let us turn to the schema of evolutionary generalism. According to D'Arms et al., evolutionary generalists "seek to describe behavior by pointing to adaptive advantages for those who engage in it, without attempting to explain how exactly tendencies to behave in the relevant way are embodied in a psychology." (D'Arms et al. 1998, 87) In order for the comparison between evolutionary generalism and evolutionary particularism to be interesting, these two explanatory schemas must address the same explanandum. Thus, as for evolutionary particularism, the explanandum for evolutionary generalism consists of a behavior *b* in a situation *S* in response to problem *p*. Although D'Arms et al. describe the generalist approach as one which does not attempt to explain how tendencies to behave in the relevant way are embodied in a psychology, we must also remember that here, as before, D'Arms et al. are talking about an instance of the explanatory schema of evolutionary generalism. What form does the explanans take when we ascend to the level of the schema? Given the remark that "what the generalist approach to evolutionary explanation lacks in detail, it seeks to compensate for with robustness," (D'Arms et al. 1998, 87) it would seem not too far off the mark to take the explanans of evolutionary generalism as a specification of some robustness properties *R* (or adaptive advantages) possessed by the behavior *b*.

As mentioned, D'Arms et al. argue that Skyrms's introduction of correlation into his replicator dynamic model of the Nash bargaining involves an "unhealthy amalgam" of the two explanatory schemas. Now that we have (hopefully) clarified what these two schemas are, the time has arrived to take a closer look at this claim.

Consider the claim that evolutionary game theoretic accounts of moral norms use the explanatory schema of evolutionary generalism. Evolutionary generalists seek to explain individual behavior by appealing to adaptive advantages accruing to individuals who engage in such behavior, without providing an explicit account of the proximate mechanisms (psychological or biological) that generate the behavior. In this eschewal of explicit details regarding proximate mechanisms, generalist explanations do stand in stark contrast with particularist explanations. Particularists attempt to account for human behavior through particular psychological (or biological) mechanisms, where these mechanisms were acquired over time by natural selection because of the particular solution they offered to the particular adaptive problems faced by individuals, and generalists do not.

An evolutionary particularist may offer the following criticism of the evolutionary generalist program: too many details are omitted for what generalists offer to count as an explanation. Although problems exist with competing evolutionary accounts of principles of justice, at least socio-

biologists and selfish-gene theorists offer possible accounts of how certain behaviors might be brought about by selective forces. Without some sort of fine-grain story of the proximate mechanisms serving to bring about the behavior in question,[2] the evolutionary generalist's appeal to selective forces serves as a naturalistic god of the gaps.[3]

Although this criticism has some force, it seems to miss the point of generalist explanations. Generalist explanations, by their very nature, do not seek the precise mechanisms underlying individual behavior. What generalist explanations provide, as I see it, is an explanation for why such behavior was selected for in the first place. This is why evolutionary generalists concern themselves with questions about the robustness properties of the behavior *b* under inspection. If one can show that behavior *b* confers adaptive benefits in all, or almost all, situations in which an agent might find herself, then selective forces will generally tend to increase the prevalence of *b* in the population.[4] If the model reasonably approximates the relevant features of the real world, we have an explanation of why one would expect to find *b* widely followed by agents in the population. It is a curious fact that the two explanatory schemas, though they address the same explanandum, operate on very different levels: particularist explanations show how certain specific mechanisms (generated by natural selection) serve to produce *b*; generalist explanations show why we might expect *b* to be selected for in the first place.

As I mentioned earlier, D'Arms et al. perceive a tension between the explanatory strategies of evolutionary generalism and evolutionary particularism. They dismiss one possible justification for introducing correlation between strategies in Skyrms's model on the grounds that it "would involve an uncomfortable amalgam of generalist and particularist explanatory schemas." Apparently, D'Arms et al. think Skyrms needs to explain what allows him to add correlation into the model without it appearing ad hoc and, furthermore, they also think that any such explanation must

2. Such as a specification of biological, psychological, or sociological mechanisms which create an increased tendency to produce the behavior, as well as a plausible account of how those mechanisms may arise as a product of selective forces on the population.

3. A weaker criticism simply notes that, " . . . when such explanations undermine our own understanding of our practices, it is appropriate to request an account of how facts about fitness have impinged themselves on the agent. Failure to provide such an account is not a decisive objection to the explanation, but . . . can often be counted against it." (D'Arms et al. 1998, 83)

4. The assumption that agents tend to adopt behaviors conferring adaptive benefits appears under various names depending on the discipline. In economics, saying that a behavior confers "adaptive benefits" is often elliptical for saying that behavior satisfies individual preferences (or increases the likelihood of satisfying individual preferences) of the agent, and subsumed under the rational actor hypothesis.

refer to some particular proximate mechanism, in violation of the explanatory generalist schema (outlined above) which does not make reference to any specific proximate mechanism.

This criticism does not seem right for two reasons. First, why should Skyrms, or any evolutionary generalist, have to appeal to specific proximate mechanisms to justify the introduction of correlation into his model? *Both* D'Arms et al. and Skyrms make certain assumptions about the nature of the underlying population when creating their models. If Skyrms needs to explain what allows him to add correlation to the model, by appealing to some particular proximate mechanism, then it seems that D'Arms et al. would similarly need to provide the particular proximate mechanisms that justify the assumptions underlying their model. Or, putting the point another way, on what grounds do D'Arms et al. identify some assumptions as needing justification in terms of proximate mechanisms while other assumptions (to my mind, equally in need of justification) get included for free? Requiring that one provide evidence of an underlying proximate mechanism supporting a choice in model construction suggests that the explanatory schema of evolutionary generalism cannot stand apart from the schema of evolutionary particularism, a position at odds with D'Arms et al.'s portrayal of generalism and particularism as distinct, independent explanatory schemas.

Second, why does an appeal to a specific proximate mechanism by an evolutionary generalist produce an "unhealthy amalgam" at all? It seems that D'Arms et al. take the generalist's intent to point to adaptive advantages (or robustness properties) of a behavior "without attempting to explain how exactly tendencies to behave in the relevant way are embodied in a psychology" as *requiring* generalist explanations to eschew referencing any specific proximate mechanism. So much so that, if a generalist does, in fact, appeal to specific proximate mechanisms, the explanation offered ceases to comply with the explanatory generalist schema.

I suspect the real reason underlying the D'Arms et al. requirement that the generalist eschew talk of specific proximate mechanisms has to do with how the evolutionary generalist shows that the behavior of interest possesses the "right sort" of sufficiently strong robustness property. The fewer specific assumptions the evolutionary generalist need appeal to in his or her game theoretic model (e.g., specific proximate mechanisms which generate correlation between certain behavior types in the population of interest), the wider range of applicability the resulting robustness property has. For example, if an evolutionary generalist can show that fair behavior emerges in a replicator dynamic model of the Nash bargaining game with a small amount of (positive) correlation, we might expect to find such behavior in situations sufficiently close to the Nash bargaining game among all species with resources (biological or psychological) capable of

generating such correlation.[5] However, if an evolutionary generalist could show that fair behavior emerges in a replicator dynamic model of the Nash bargaining game without any correlation (which we know is not possible), then we would expect to find such behavior present in a wider range of species.

So, when evolutionary generalists do appeal to specific proximate mechanisms to justify some aspect of a model, that serves to reduce the strength of the resulting robustness claim. However, we should recognize that the amount which appeal to specific proximate mechanisms reduces the resulting robustness claim depends entirely on the nature of the proximate mechanism appealed to. To be sure, if a generalist was only able to show (in an appropriately formed game-theoretic model) that fair division in the Nash bargaining game occurred in populations of philosophers who all correlated on the strategy followed by the one named Brian, then the resulting robustness claim (and the consequent strength of the explanation) would be very weak indeed. Luckily, most cases of interest will not be this extreme.

Setting this issue aside, I now argue for the connection between the schemas of generalism and particularism being considerably closer than D'Arms et al. allow. As we just saw, requiring the generalist to eschew proximate mechanisms entirely would force the generalist to work with such abstract models that it might be difficult, in principle, to justify including processes required to model the desired phenomena accurately. Constructing good evolutionary game theoretic models involves the delicate task of choosing which features to include and which to neglect, a task virtually impossible to do well without paying considerable attention to the details that D'Arms et al. envision the evolutionary generalist sweeping under the rug. Consequently, I will sketch a two-tiered strategy of evolutionary explanation in which the generalist and particularist schemas appear as conceptually distinct, but necessarily integrated, components.

Consider the general problem at hand: to what extent can we give an evolutionary explanation of human behavior? Before we can make much progress on this question, we obviously need to narrow the scope of the question through further specification of the explanandum. Recently, a favored explanandum has been the existence of altruistic behavior. As is well known, this was the topic initially chosen by sociobiologists and selfish gene theorists because, on the surface, it seems that Darwinian natural selection should exclude the emergence of altruism. (See Sober and Wilson

5. I qualify this somewhat since other factors of greater import to the survival of the species might trump the adoption of fair division in Nash-bargaining-game-like situations.

(1998) for an extended argument that Darwinian natural selection can favor altruistic behavior when properly understood as multilevel selection.) Skyrms's work on fair division in games of divide-the-dollar chooses a different explanandum, a behavior widely observed in both informal and formal settings (see Nydegger and Owen 1974; Huyck et al. 1995; Yaari and Bar-Hillel 1984).

Once the explanandum has been selected, we need to determine the level at which we seek an explanation. Suppose that, as good reductionists, we seek a finely-structured explanation for the behavior b in question where, ideally, this means specifying some mechanism underwriting b. Looking for such an explanation, we approach the explanatory question as an evolutionary particularist. Although *some* mechanism has to exist, simply because any naturally occurring behavior has some mechanism which brings it about, what we, as evolutionary particularists, are interested in is whether there is an *adaptive* mechanism underwriting the behavior. Yet one would be ill advised to search for an adaptive mechanism without having reason to believe an adaptive mechanism exists. After all, b could be generated as a side effect from two (or more) mechanisms operating concurrently, each of which was selected for reasons having little to do with the consequences of b.[6]

However, if one can show that behavior b emerges from all, or almost all, initial situations in a model M that reasonably approximates the relevant features of the situation under consideration, then we have good reason to believe that an adaptive mechanism underwriting b exists. The general principle here is that if a behavior confers a strong selective advantage to individuals who follow that behavior, then selective forces (cultural or biological) will operate so as to install proximate mechanisms (again, cultural or biological) which, when enacted, realize behavior b. At this point, we now have reason to believe that an adaptive mechanism exists.[7]

In this two-tiered conception of evolutionary explanation, one uses abstract, idealized models which capture a sufficient level of detail of the

6. This differs from the problem of functional equivalents. The problem of functional equivalents notes that if an organism evidences behavior b in environment e, and e is correlated with another property i, then b may be explained by either a proximate mechanism p detecting e or a proximate mechanism p' detecting i. That is, the process of natural selection need not choose proximate mechanisms for the behavior b which take into account the environment directly. The possibility noted here is that if the organism has proximate mechanisms p and p' operating concurrently, then the interaction of the two mechanisms may generate b when p and p' were selected for reasons which have nothing to do with the adaptive benefits conferred by b.

7. I do not mean to attribute this methodology to either evolutionary game theory or Skyrms. It is merely put forth as a recommendation.

actual situation to determine which behaviors are likely candidates for being generated by adaptive mechanisms. Once we have established the plausibility of there being an underlying adaptive mechanism, to complete the account we then look for a finely-structured explanation given in terms of proximate mechanisms. However, as the question of what proximate mechanisms give rise to behavior is an empirical question, it is one best addressed by wet biologists, sociologists, anthropologists, cognitive psychologists, and so on.

The lower tier, then, attempts to explain the substantive content of principles of justice in terms of particular adaptive mechanisms. If we cannot discern a mechanism, brought about by social or biological evolution, which generates the behavior, something is missing in the explanation. Additionally, the second tier is needed because no matter how successful we are at explaining the substantive content of a principle of justice in terms of particular adaptive mechanisms, we still need to account for the *normative* content of the principle.[8] This is, I believe, why the generalist approach plays an essential role, and why evolutionary explanations of justice can not be given entirely in particularist terms.

We have already supposed that the generalist has shown, in a model M which reasonably approximates the relevant features of the situation under consideration, that from all (or almost all) initial conditions the population converges to a state in which the behavior b dominates. If M did not provide for the random introduction of new strategies into the population (say, via mutation or trembling-hand type errors), let M' be a model extending M which does. (If M already considered mutations or trembling-hand type errors, then $M' = M$.) Finally, suppose that for reasonable values of the mutation parameters $\hat{\mu}$ one can show that it still holds that from all (or almost all) initial conditions the population converges to a state in which the behavior of interest dominates. We also need to add the requirement that in the unlikely event every member (or most members) of the population mutates into a strategy other than fair divi-

8. If we have a rich enough set of proximate mechanisms, we might be able to explain the *perceived* normative content of the principle. For example, someday we might be able to explain why we feel that the 50-50 split in divide-the-dollar is the right thing to do because of various features of the complex neurological architecture of the brain, coupled with our particular learning histories, combined with a description of how the process of natural selection led to our present neurological architecture. However, one can always ask whether the 50-50 split in divide-the-dollar is *really* the right thing to do, regardless of how we feel about it. It is conceivable that, someday, we might also be able to explain the proximate mechanisms (neurological or sociological) generating this metaphysical questioning in such a way so as to render the question meaningless. However, let us assume for the time being that this question is meaningful and that what one is asking for is an explanation of why one *ought* to demand 50-50 in the game of divide-the-dollar.

sion, fair division will shortly dominate again. Let us call a strategy with this property a *stochastically robust strategy*.[9] If such conditions hold, it seems to me that this would provide an explanation for the *normative* component of the behavior.

To see why, consider the case for fair division in divide-the-dollar. What does it mean to say that, under completely symmetric circumstances, one *ought* to demand half? This can mean several things. It can mean that it is morally wrong (or unjust or unfair) to do otherwise, or, alternatively, that it is simply in one's self-interest to demand half. Although both cases say that one should demand half, some see the former as having an additional normative component that the latter lacks. Here, I concentrate exclusively on self-interest, assuming the purported extra moral content to be a useful illusion.[10]

If fair division were a stochastically robust strategy, would an agent A have reason to follow a strategy other than fair division? If the population consisted mostly of fair dividers, A's adopting a strategy other than fair division would correspond to a mutation occurring in the population. Since we are assuming fair division to be stochastically robust, the majority of the population will continue to follow fair division even in light of A's mutation. Furthermore, adopting another strategy other than fair division does not work to A's advantage: if he demands more than half of the cake, he will receive nothing in his interactions with fair dividers and, since they constitute the majority of the population, this means that in most of his interactions A will receive nothing. On the other hand, if A demands less than half the cake, in the majority of his interactions he will receive less than he would have if he demanded half. In such a population,

9. This last requirement is not the same as requiring the strategy of fair division to be stochastically stable. Recall that if one can show the amount of time the population spends in the state of fair division converges to one as the mutation rate goes to zero, the strategy producing the behavior of concern is a stochastically stable strategy. The reason for the stronger requirement shall be explained shortly.

10. A complete treatment of this point would take us too far afield, but a few remarks are called for. Consider the utilitarian's response to the charge that they do not really account for the moral sentiment behind norms: since agents cannot perform the utility calculations justifying acceptance of the rule, we endow the rule with supposed "moral" force to ensure compliance. The parallel problem here is that boundedly rational agents cannot reproduce the generalist's argument that demanding half best serves their self-interest, but they can know *that* it is in their self-interest to demand half. Thus, these boundedly rational agents endow the rule of demanding half with supposedly "moral" force to ensure compliance. Agents can know that it is in their self-interest to demand half, without knowing why, by simple induction: each agent keeps track of how well her surrounding neighbors do over time, discovering that agents who demand half typically do better than agents who do not. Although no agent knows *why* demanding half typically does better than not, all can detect that they will do better if they demand half than if they do not.

A should demand half because *only* the strategy of demand half maximizes A's expected amount of cake.

Now consider what happens when the population consists of strategies other than fair dividers, such as one of the polymorphisms occurring in the models of Skyrms (1996) and D'Arms et al. (1998). Would A have reason to follow a strategy other than fair division? Given the definition of a stochastically robust strategy, fair division will shortly come to dominate the population even though the population does not currently contain any fair dividers. A might receive limited benefits from continuing to follow a strategy belonging to the polymorphic pair, but once the majority of the population has switched to fair division we are back in the case discussed above. If A seeks to maximize his expected amount of cake, A will ultimately adopt the strategy of demand half. Since fair division will shortly dominate the population, A should demand half because only the strategy of demand half maximizes A's expected amount of cake.

I am misspeaking slightly in saying that the strategy of demand half maximizes A's expected amount of cake, since I have not said anything about the subjective probabilities that A assigns to the strategies of his opponents. This is deliberate, and depends on a particular conception of the strategic problem A faces. I assume that A is a boundedly rational agent who only has knowledge of the immediate players he interacts with, where those neighbors constitute a very small segment of the total population. (This conception underlies the spatial model presented in Section 4.) In such a situation, we do not need to speak of the subjective probabilities A assigns to his opponents' strategies since A *knows* his opponents' strategies.

However, since A has no information about the strategies of his opponents' opponents, A can infer nothing about the future strategies of his opponents. Why? Because the future strategy of A's opponent depends upon the strategies held by the opponents of A's opponent.[11] Although A does know the current strategy of his opponents, this is of *no use* to him when deciding whether he should change his strategy during the next generation. In other words, A's choice of strategy for the next generation will be a decision under uncertainty, not a decision under risk.

Yet if demand half is a stochastically robust strategy and A knows this, A knows that the strategy of fair division will shortly dominate the population. This provides a strong incentive for A to demand half, since, given A's expectation that fair division will shortly dominate the population,

11. Strictly speaking, A does have knowledge of one of his opponents' opponents, since A knows his own strategy. Assuming, though, that each player has a sufficiently large number of opponents, this does not give A enough information to infer anything about his opponents' future strategies.

demand half is the strategy which will maximize *A*'s amount of cake when fair division comes to dominate. Since *A* has no information as to *when*, exactly, fair division will dominate (other than that it will shortly), if *A* adopts any strategy other than fair division, *A* will not be acting to maximize expected utility.

One might be tempted to argue that *A*, rather than adopting the strategy of fair division, should instead adopt a best-response strategy (taking into consideration his opponents' strategies). However, there are good reasons for thinking that *A* should *not* do this. A best-response strategy for *A* will be determined according to the strategies *A*'s neighbors currently hold. However, in many cases some (or all) of *A*'s neighbors will change strategies in the next generation, often rendering *A*'s best-response strategy less effective than fair division (see Alexander, 2000, for an example of this effect). Thus, what the excursion through stochastic robustness, and its dependence on the future trajectory of the population, provides us with is one possible motivation for agents adopting the strategy of fair division.

Notice the requirement that fair division be a stochastically robust strategy cannot be replaced by weaker assumptions, for instance that it be an evolutionarily stable strategy or a stochastically stable strategy. Fair division is the unique evolutionarily stable strategy of divide-the-dollar, but that does not mean a single agent in a $\frac{1}{3}$–$\frac{2}{3}$ polymorphic population should consider adopting it. As the models of Skyrms and D'Arms et al. show, a $\frac{1}{3}$–$\frac{2}{3}$ polymorphic population can resist invasion by fair dividers to a considerable extent. Neither demand $\frac{1}{3}$ nor demand $\frac{2}{3}$ are, by themselves, evolutionarily stable, but the polymorphism containing both resists invasion quite well.

The concept of a stochastically stable strategy does not fit the bill either, since it places no requirement on the amount of time it takes for the population to move out of an unfair polymorphism. Fair division is the unique stochastically stable strategy in the game of divide-the-dollar, but this does not mean a single agent in a $\frac{1}{3}$–$\frac{2}{3}$ polymorphic population should consider adopting it. All it means to say that a strategy is stochastically stable is that, in the limit as the mutation rate converges to 0, the proportion of time the population spends in the pure state of fair division converges to one—there is no mention about how rapidly the population moves out of an unfair polymorphic state. It may very well be in the best interests of an agent trapped in a $\frac{1}{3}$–$\frac{2}{3}$ polymorphism to continue with a strategy of demand $\frac{1}{3}$ or demand $\frac{2}{3}$, if the population will remain in that polymorphism for the next several hundred (or thousand) generations.

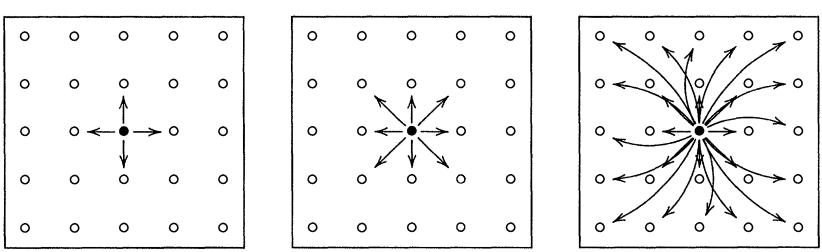
At this point, we have no reason to believe that fair division in the game of divide-the-dollar is stochastically robust. The next section develops a spatial version of the game of divide-the-dollar, one more realistic than the model of Skyrms (1996) but differing considerably from the model of

D'Arms et al. (1998). This spatial model has the property that fair division dominates in almost all cases when mutations are not present, and dominates in all cases when mutations are present. I take this to signify two things: first, it suggests the strategy of fair division is stochastically robust; second, it demonstrates that the results D'Arms et al. claimed to undermine Skyrms's account of the evolution of justice depend entirely on the model used. (A proof that fair division is, in fact, a stochastically stable strategy can be found in Alexander (1999).)

4. A Spatial Game Of Divide-The-Dollar. In this model we consider a finite population of agents distributed over a rectangular lattice that does not connect at the edges. The *neighborhood* of a player p, denoted $N(p)$, is the set of all players q that p interacts with during a given round of play. A round of play consists of two stages. In the first stage, a player p plays the game of divide-the-dollar with every player in his neighborhood, earning a score equal to the sum of payoffs from each individual game. In the second stage, the player updates his strategy by comparing his success level with that of every player in his neighborhood. Although no *a priori* reason exists for assuming these neighborhoods to be equal, I follow the majority of papers in the spatial modeling literature by assuming they are.

Commonly studied neighborhood types when the underlying structure of the world is a rectangular lattice are listed in Figure 1. These diagrams specify directional offsets identifying the neighbors of a player (indicated in the diagram by a filled circle). Since the models of this paper are bounded, players on the boundary have fewer neighbors than those in the interior.

Given that the models of Skyrms (1996) and D'Arms et al. (1998) do not consider a finite set of agents positioned on a lattice, the connection between this model and the Skyrms-D'Arms et al. debate may not be



(a) von Neumann (b) Moore (8) (c) Moore (24)

Figure 1. Three common neighborhoods defined on a square lattice.

immediately clear. Recall D'Arms et al.'s criticism of the introduction of correlation into Skyrms's replicator dynamic model:

> What is the justification for adding a correlation factor, though? Once Skyrms relaxes the requirement of random interactions in the population, and allows some degree of assortative interactions, we need to hear a justification for assuming that the likely departure from random interactions will be toward correlation in particular. Why think that individuals are especially likely to meet others playing the same strategy as they play? (D'Arms et al. 1998, 92)

D'Arms et al. suggest that one likely justification for the introduction of correlation would be a scenario in which individual strategies were influenced by genes. The spatial model described here demonstrates one-way correlation between strategies can arise without positing a genetic influence. Here, correlation between strategies emerges simply through their spatial positioning and the fact that individuals interact only with their immediate neighbors.

One might also be under the impression that individual evolutionary game theoretic accounts of the evolution of justice require a commitment to a particular model or class of models. If so, then the spatial model developed below will seem irrelevant to the current debate. This, I believe, mistakenly takes one's commitment to the project of seeking evolutionary grounds for certain moral concepts as a commitment to a particular model used to illustrate how such evolutionary grounds might arise. Skyrms's remarks in *Evolution of the Social Contract* caution against such a move:

> In a finite population, where there is some random element in evolution, some reasonable amount of divisibility of the good and some correlation, we can say that it is likely that something close to share and share alike should evolve in dividing-the-cake situations. This is, perhaps, a beginning of an explanation of the origin of our concept of justice. (Skyrms 1996, 21)

Since the replicator dynamic model is only the beginning of the explanatory story, we should not think that evolutionary game theoretic accounts of justice depend upon it.

4.1 Dynamics. The model considered here allows for three different update rules, each rule having a certain degree of plausibility. The general question of how one's choice of the update rule affects the limit form of the model remains an open and difficult problem.

Imitate the best neighbor. This is the most common update rule in the spatial modeling literature (for some instances, see Nowak and May 1992; Nowak and May 1993; Lindgren and Nordahl 1994; Huberman and

Glance 1993; Epstein 1998). According to this rule, each player p looks at her neighbors and adopts the strategy of the neighbor who did the best, where "best" means "earned the highest score." As ties may occur between several players in the neighborhood of p, an additional rule needs to be given which, in such circumstances, selects a unique strategy.[12] (In all cases it is assumed that p does not change her strategy unless one neighbor did strictly better than her.)

Imitate with probability proportional to success. Unlike the previous update rule, which ignored neighbors who did better than p but did not earn one of the best scores, this rule assigns to every neighbor q who did better than p a nonzero probability that p will adopt q's strategy. The exact probability that p will adopt q's strategy increases linearly with the relative success of q (for more details on this rule and the others, see Alexander 1999).

Imitate best average payoff. Under these dynamics, players calculate the average payoff of each strategy in their neighborhood and select the one with the highest value. Since the possibility of ties exists, as in the case of imitate the best neighbor, some kind of tie-breaking rule needs to be given. Formally, the tie-breaking rule is the same as that for imitate the best neighbor, with the exception that we use the set of all strategies which tied for the title of "best average payoff" instead of the set of strategies which earned the highest score.

All three update rules assume some sort of imitation dynamic in which players mimic those who did "best" according to some criteria. This deviates somewhat from the standard game theoretic tradition, which typically assumes that players employ more strategic update rules, such as adopting a best-response strategy or seeking out compatible players to interact with. The use of imitation rules fits better with the assumption that agents are only boundedly rational and tend to follow reliable heuristics instead of expressly calculating the optimal response in each situation.

12. Call a strategy which earned one of the highest scores in the neighborhood of p a *maximal* strategy. We assume that the number of players in $N(p)$ who follow a given maximal strategy s affects the likelihood that p will choose to adopt s. This seems reasonable since, if several neighbors of p follow s and earn the maximal score of $N(p)$, it would be foolish of p to ignore this information. One simple way p might take this information into account is to let the probability of choosing a maximal strategy s be a linear function of the number of people in $N(p)$ who follow that strategy. (More complicated functions could be used to model risk-averse players who require a certain number of neighbors to follow a maximal strategy before they consider adopting it.) For simplicity, we assume that if the number of players in $N(p)$ using maximal strategy s is n_s, then the probability of p choosing to adopt s is n_s divided by the total number of neighbors who earned the highest score.

4.2 Synchronicity Assumptions. We assume all updating occurs synchronously. There has been considerable debate over the appropriateness of this assumption. The original papers of Nowak and May (1992, 1993) on the spatialized prisoner's dilemma used synchronous dynamics, and were later criticized by Huberman and Glance (1993) on the grounds that synchronous dynamics lead to stable equilibrium states which did not appear when asynchronous dynamics were used. Since then, asynchronous dynamics have typically been preferred, as the more recent papers of Hegselmann (1996) and Epstein (1998) indicate. I do not believe, though, that asynchronous dynamics necessarily offer a more accurate model as they are usually purported; although agents do not update their strategies in the rigid lock-step manner suggested by synchronous dynamics, neither do they update their strategies in the carefully orchestrated manner of asynchronous dynamics, where only one agent changes her strategy at a time.

4.3 Results. Table 1 summarizes the final convergent state of the world for several different combinations of neighborhoods and dynamics. The neighborhoods examined include the three most common in the literature (von Neumann, Moore (8), and Moore (24)), as well as the three nonstandard types displayed in Figure 2. The row identified as "R(8)" used a different method: at the start of every generation, each player p randomly selects eight players from the world to serve as p's neighborhood for interaction and updating. Thus, the model of row R(8) serves as an intermediary between the fixed neighborhood structure of this model and models based on the replicator dynamics.

In general, mean times to convergence are quite rapid, as Table 2 shows. Models using the Moore (8) neighborhood usually converged within sixteen generations to fair division. This is a considerable improvement over the results of Skyrms (1996), and a significant improvement over that of Kandori et al. (1993), whose stochastically stable equilibrium only selects the equilibrium of fair division in the limit. The larger Moore (24) neighborhood leads to faster convergence times because the radius of influence of any given single player has increased.

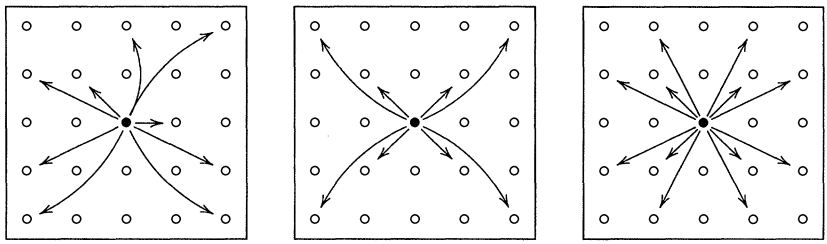
Figures 3, 4, and 5 illustrate the evolutionary path followed by worlds using three nonstandard neighborhoods. In these figures, the initial conditions set all strategies equally likely and had players update their strategies using imitate the best neighbor dynamics. In the first two worlds, the strategy of fair division emerges from the initial random conditions in the absence of means to globally coordinate such an outcome. The third figure illustrates the effect of a degenerate (one-person) neighborhood in which all players use only their northern neighbor for interaction and updating.

TABLE 1: Convergence results based on neighborhood and dynamic.

Nbhd	Dynamics	Polymorphism						
		0–10	1–9	2–8	3–7	4–6	5	Other
VN	Mimic with proportion relative to success	0	0	0	0	29	9970	1
	Mimic best neighbor	0	0	0	0	26	9966	8
	Imitate best average strategy	0	0	0	0	13	9984	3
M(8)	Mimic with proportion relative to success	0	0	0	0	26	9973	1
	Mimic best neighbor	0	0	0	0	26	9908	66
	Imitate best average strategy	0	0	0	0	24	9970	6
M(24)	Mimic with proportion relative to success	0	0	0	8	110	9879	3
	Mimic best neighbor	0	0	0	21	220	9721	38
	Imitate best average strategy	0	0	0	0	62	9934	4
R(8)	Mimic with proportion relative to success	0	0	57	556	2418	6964	5
	Mimic best neighbor	0	0	54	550	2560	6833	3
	Imitate best average strategy	0	0	0	1	1523	8439	37
Type 1	Mimic with proportion relative to success	0	0	0	3	47	9949	1
	Mimic best neighbor	0	0	0	3	62	9933	2
	Imitate best average strategy	0	0	0	0	29	9962	9
Type 2	Mimic with proportion relative to success	0	0	0	0	32	9899	69
	Mimic best neighbor	0	0	0	0	43	9868	89
	Imitate best average strategy	0	0	0	0	28	9924	48
Type 3	Mimic with proportion relative to success	0	0	0	3	42	9950	5
	Mimic best neighbor	0	0	0	3	62	9933	5
	Imitate best average strategy	0	0	0	0	32	9965	3

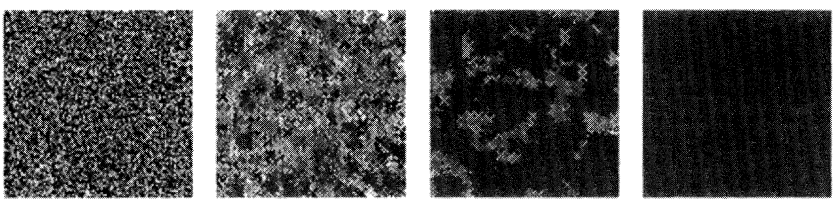
TABLE 2: Mean convergence times

Nbhd	Dynamics	Population composition					
		0–10	1–9	2–8	3–7	4–6	fair
VN	Mimic with proportion relative to success	—	—	—	—	25.4	3.9
	Mimic best neighbor	—	—	—	—	22.7	26.3
	Imitate best average strategy	—	—	—	—	32.2	23.9
M(8)	Mimic with proportion relative to success	—	—	—	—	17.9	16.4
	Mimic best neighbor	—	—	—	—	28.0	15.4
	Imitate best average strategy	—	—	—	—	17.2	14.6
M(24)	Mimic with proportion relative to success	—	—	—	13.9	17.2	14.9
	Mimic best neighbor	—	—	—	32.3	22.8	12.8
	Imitate best average strategy	—	—	—	—	18.7	10.6
R(8)	Mimic with proportion relative to success	—	—	38.4	24.8	13.5	6.2
	Mimic best neighbor	—	—	15.5	13.9	8.5	4.5
	Imitate best average strategy	—	—	—	28.0	16.1	5.37
Type 1	Mimic with proportion relative to success	—	—	—	22.0	21.8	15.2
	Mimic best neighbor	—	—	—	24.0	10.7	12.69
	Imitate best average strategy	—	—	—	—	11.2	11.4
Type 2	Mimic with proportion relative to success	—	—	—	—	22.4	15.6
	Mimic best neighbor	—	—	—	—	16.3	15.3
	Imitate best average strategy	—	—	—	—	16.6	13.7
Type 3	Mimic with proportion relative to success	—	—	—	17.7	20.7	14.4
	Mimic best neighbor	—	—	—	24.0	10.7	12.7
	Imitate best average strategy	—	—	—	—	11.5	12.0



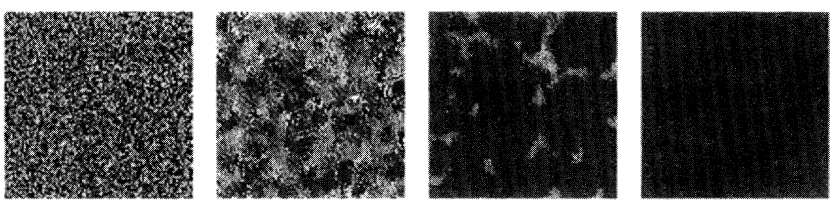
(a) Type 1 (b) Type 2 (c) Type 3

Figure 2. Three nonstandard neighborhoods used on Table 1.



(a) Gen. 0 (b) Gen. 2 (c) Gen. 4 (d) Gen. 10

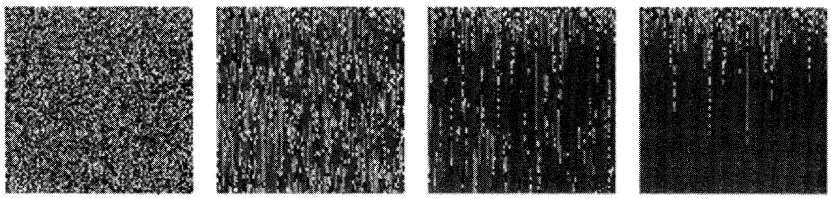
Figure 3. Evolution under neighborhoods of type 1.



(a) Gen. 0 (b) Gen. 2 (c) Gen. 4 (d) Gen. 6

Figure 4. Evolution under neighborhoods of type 2.

4.4 Dependence Upon Cake Size. Skyrms (1996) reported an interesting relationship between granularity of the good and the distribution of the resulting polymorphism. It turns out that increasing the total number of pieces into which the cake is sliced leads to an increase in the total number of populations that will evolve into something "near" fair division. In particular, Skyrms found that a cake divided into 200 pieces went to fair division ± 3 pieces approximately 94.1% of the time; all trials went to fair



(a) Gen. 0 (b) Gen. 10 (c) Gen. 30 (d) Gen. 100

Figure 5. Evolution under degenerate (one person) neighborhoods.

division ± 11 pieces. Since most populations evolving under spatial constraints already lead to a pure state of fair division, the natural question in this context becomes how coarse can we slice the cake while still getting fair division? Table 3 lists the results as the number of slices varies from ten to two, for each of the three dynamics considered, under the Moore (8) neighborhood.

4.5 Mutations. Mutations introduce a small amount of stochasticity into the model, controlled by a single global mutation rate μ. At the end of each generation, each individual in the population has probability μ of adopting another strategy. Since the probability of no mutations occurring during a single generation is quite low, even for relatively small populations, we must adjust our concept of convergence accordingly. I shall say that a population converges to a state where fair division dominates if all but $N \cdot \mu$ members of the population follow the strategy of fair division (where N denotes the total size of the population). In a series of 10,000 trials (all beginning from a randomly chosen point in state space) with a mutation rate $\mu = .001$, all trials converged to a state where fair division dominated. Figure 6 illustrates how a pure 4–6 polymorphism may be taken over by fair division in the presence of a little mutation.

The amount of time required to move a population out of a polymorphism to a state where fair division dominates obviously depends on the frequency of mutations μ. Inspection of Figure 6 reveals that the critical step involves the introduction of the demand ½ strategy into a site surrounded by sufficiently many compatible strategies. If μ is large, we do not have to wait very long for such a mutation to occur. If μ is small (or if there are not many sites following strategies compatible with fair division), longer times are required. However, even when μ is small the total time required is quite small in comparison with the time for the model of Kandori et al. (1993).

TABLE 3: Convergence results for a varying number of slices[1]

Cake size	Dynamics	Polymorphism						
		0–10	1–9	2–8	3–7	4–6	5	Other
10	Mimic best neighbor	0	0	0	0	2	998	0
	Imitate best average strategy	0	0	0	0	2	998	0
	Imitate using relative success	0	0	0	0	3	997	0
9	Mimic best neighbor	0	0	0	0	5[10]	0	995[11]
	Imitate best average strategy	0	0	0	0	1*	0	999[12]
	Imitate using relative success	0	0	0	0	17*	0	983[13]
8	Mimic best neighbor	0	0	0	0	999[1]	0	1[2]
	Imitate best average strategy	0	0	0	0	998*	0	2[3]
	Imitate using relative success	0	0	0	0	1000*	0	0
7	Mimic best neighbor	0	0	0	3	0	0	997[14]
	Imitate best average strategy	0	0	0	3	0	0	997[15]
	Imitate using relative success	0	0	0	4	0	0	996[16]
6	Mimic best neighbor	0	0	0	998*	0	0	2[4]
	Imitate best average strategy	0	0	0	1000*	0	0	0
	Imitate using relative success	0	0	0	995*	0	0	5[5]
5	Mimic best neighbor	0	0	1	0	0	0	999[17]
	Imitate best average strategy	0	0	1	0	0	0	999[18]
	Imitate using relative success	0	0	2	0	0	0	998[19]
4	Mimic best neighbor	0	0	1000*	0	0	0	0
	Imitate best average strategy	0	0	999*	0	0	0	1[6]
	Imitate using relative success	0	0	997*	0	0	0	3[7]
3	Mimic best neighbor	0	1	0	0	0	0	999[20]
	Imitate best average strategy	0	0	0	0	0	0	1000[21]
	Imitate using relative success	0	0	0	0	0	0	1000[22]
2	Mimic best neighbor	0	997*	0	0	0	0	3[8]
	Imitate best average strategy	0	1000*	0	0	0	0	0
	Imitate using relative success	0	998*	0	0	0	0	2[9]

1. [1]Of these, 973 were pure states of demand 4. [2]A 3–5 polymorphism, (3609, 6391). [3]Two 3–5 polymorphisms: (5112, 4888), (5196, 4804). [4]Two 2–4 polymorphisms: (3736, 6264), (3813, 6187). [5]Five 2–4 polymorphisms: (3273, 6727), (3484, 6516), (3380, 6620), (3476, 6524), (3589, 6411). [6]A 1–3 polymorphism, (2563, 7437). [7]Three 1–3 polymorphisms: (2147, 7853), (2233, 7767), (2135, 7865). [8]In all three worlds, the strategy of demand 1 went extinct early on, leaving the population in an unstable equilibrium of (1,0,9728,34,34,36,1,90,28,0,48), (22, 0, 5682, 129, 571, 884, 556, 430, 960, 128, 638), and (69, 0, 5771, 1646, 321, 86, 187, 626, 844, 115, 335). [9]Both worlds contain unstable equilibrium in which all strategies are present: (4, 7, 5241, 212, 90, 195, 280, 495, 2387, 572, 517) and (2, 11, 2440, 523, 646, 988, 702, 1831, 105, 1040, 1712). [10]Four of these states contained only demand 4. [11]One 3–6 polymorphism, one 3–5 polymorphism, with the rest being 4–5 polymorphisms. [12]Three 3–6 polymorphisms, the rest 4–5 polymorphisms. [13]Three 3–6 polymorphisms, the rest 4–5 polymorphisms. [14]One 2–5 polymorphism, one 2–4–5 polymorphism, the rest 3–4 polymorphisms. [15]Five 2–5 polymorphisms, the rest 3–4 polymorphisms. [16]Two 2–5 polymorphisms, the rest 3–4 polymorphisms. [17]Two 1–3–4 polymorphisms, the rest 2–3 polymorphisms. [18]One 1–3–4 polymorphism, the rest 2–3 polymorphisms. [19]All 2–3 polymorphisms. [20]One world containing the unstable equilibrium (15,3,108,4220,254,146,525,855,2335,1520,19), the rest 1–2 polymorphisms. [21]One world containing the unstable equilibrium (0,0,1074,1070,1011,1111,1125,1083,1110,1175,1241), the rest 1–2 polymorphisms. [22]Three unstable equilibriums of the following form: (6,0,8764,9,10,806,60,95,116,53,81), (86,0,2958,1357,648,2611,263,1159, 618,81,219), and (10,10,129,1650,4478,4,289,253,964,1080,1133), the rest 1–2 polymorphisms.

*All states contain only the strategy making the lowest demand of the pair.

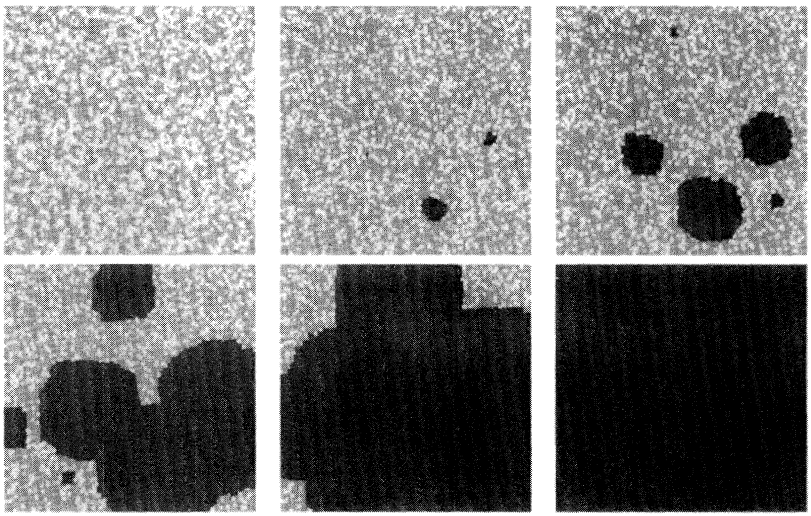


Figure 6. Emergence of fair division out of a 4-6 polymorphism due to mutation.

5. Concluding Remarks. Evolutionary accounts of justice attempt to explain principles of justice as the natural outcome of an evolutionary process operating on a population of agents. Ideally, such explanations need to account for the substantive and normative content of principles of justice. An example of the beginnings of such an explanation can be found in Skyrms (1996), which contains an evolutionary game theoretic account of how one substantive principle of justice (fair division in the game of divide-the-dollar under completely symmetric circumstances) might have come about.

D'Arms et al. (1998) criticize Skyrms's account on methodological grounds, charging that his game theoretic account violates the explanatory strategy to which he has committed himself. As argued in Section 3, this criticism depends upon an overly simplistic conception of the explanatory strategy employed. When we evaluate Skyrms's account according to the two-tier model of evolutionary explanation, the methodological criticism evaporates.

D'Arms et al. also describe results from an alternative model which, they claim, undermine Skyrms's account of the evolution of justice. One may reasonably ask *why* these results undermine his account. After all, since the replicator dynamics only capture the most elementary features of real populations, abstracting away many (possibly relevant) features that a more complete model would include, one should not take the results of this model as conclusive. Skyrms, well aware of this, consequently re-

frains from claiming that this model provides more than a first approximation of one possible process from which our concept of justice might have emerged.

Although the model of D'Arms et al. improves upon that of Skyrms, both make assumptions that limit their applicability to the evolution of justice. The spatial model presented earlier in this paper is more realistic and has more robust convergence properties than both. Unlike Skyrms's model, in the spatial model correlation between compatible strategies arises naturally through the positioning of agents: demand halfers gather around fellow demand halfers, and other polymorphic pairs gather around each other. Unlike D'Arms et al.'s model, there is no need to introduce explicit avoidance behavior between strategies: the strategies of demand $\frac{1}{3}$ and demand $\frac{2}{3}$ tend to spread out so as to minimize incompatible contact. Furthermore, in spatial models there is no need for the questionable renormalization of the population between rounds which D'Arms et al. use.

Determining which models best capture the relevant features of populations of human agents requires careful attention to nontrivial modeling issues. I do not claim that the spatial agent-based models developed here provide the best basis for the second tier of evolutionary explanations described in Section 3. However, given the extent to which the convergence properties of spatial models can differ from their replicator dynamic counterparts and the consequent new perspective offered on the evolution of the social contract, they open interesting possibilities for future research.

REFERENCES

Alexander, Jason and Brian Skyrms (1999), "Bargaining with Neighbors: Is Justice Contagious?", *Journal of Philosophy* 96, 11: 588–598.

Alexander, Jason M. (1999), "The (Spatial) Evolution of the Equal Split", Technical report, Institute for Mathematical Behavioral Sciences, U.C. Irvine.

Alexander, J. McKenzie (2000), "Artificial Justice", Forthcoming in *Artificial Life VII: Proceedings of the Seventh International Conference.* MIT Press.

Epstein, Joshua A. (1998), "Zones of Cooperation in Demographic Prisoner's Dilemma", *Complexity* 4 (2): 36–48.

Hegselmann, Rainer (1996), "Social Dilemmas in Lineland and Flatland", In Liebrand and Messick, eds., *Frontiers in Social Dilemmas Research*, Springer, pp. 337–361.

Huberman, Bernardo A. and Natalie S. Glance (1993), "Evolutionary games and computer simulations", *Proc. Natl. Acad. Sci.* 90: 7716–7718.

Huyck, John Van, Raymond Battalio, Sondip Mathur, and Patsy Van Huyck (1995), "On the Origin of Convention: Evidence from Symmetric Bargaining Games", *International Journal of Game Theory* 24: 187–212.

Kandori, Michihiro, George J. Mailath, and Rafael Rob (1993), "Learning, Mutation, and Long Run Equilibria in Games", *Econometrica* 61, (1): 29–56.

Lindgren, Kristian and Mats G. Nordahl (1994), "Evolutionary dynamics of spatial games", *Physica D* 75: 292–309.

Nash, John F. (1950), "The Bargaining Problem", *Econometrica* 18: 155–162.

Nowak, Martin A. and Robert M. May (1992), "Evolutionary games and spatial chaos", *Nature* 359: 826–829.

———. (1993), "The Spatial Dilemmas of Evolution", *International Journal of Bifurcation and Chaos* 3, 1: 35–78.

Nydegger, R. V. and G. Owen (1974), "Two-Person Bargaining: An Experimental Test of the Nash Axioms", *International Journal of Game Theory* 3, 4: 239–249.

Skyrms, Brian (1996), *Evolution of the Social Contract*, Cambridge: Cambridge University Press.

Sober, Elliot and David S. Wilson (1998), *Unto Others: The Evolution and Psychology of Unselfish Behavior*, Cambridge: Harvard University Press.

Taylor, Peter D. and Leo B. Jonker (1978), "Evolutionary Stable Strategies and Game Dynamics", *Mathematical Biosciences* 40: 145–156.

Yaari, Menachem E. and Maya Bar-Hillel (1984), "On Dividing Justly", *Social Choice and Welfare* 1: 1–24.